MoM: Multi-to-Multi Identification of Decoupled Multifaceted Clustering with Dynamic Representation Fusion

Peipei Zhang
Zhejiang-New Zealand Joint Laboratory on Vision-Based Intelligent Metrology, College of Information Engineering
China Jiliang University
Hangzhou, China
zhangpeipei@cjlu.edu.cn

Wenjie Zhu✉
Zhejiang-New Zealand Joint Laboratory on Vision-Based Intelligent Metrology, College of Information Engineering
China Jiliang University
Hangzhou, China
zhwj@cjlu.edu.cn

Qinli Zhou
Zhejiang-New Zealand Joint Laboratory on Vision-Based Intelligent Metrology, College of Information Engineering
China Jiliang University
Hangzhou, China
qinlizhou@cjlu.edu.cn

Weiqi Yan
Department of Computer and Information Sciences
Auckland University of Technology
Auckland, New Zealand
weiqi.yan@aut.ac.nz

Abstract

Multifaceted clustering is dedicated to excavating diverse and heterogeneous facets from complex multimedia data, which is a critical task for multimedia content understanding. Existing approaches primarily focus on decomposing a single dataset into multiple subspaces or sequentially generating alternative clusterings. However, these methods are severely constrained by the inherent limited information contained in the data itself, resulting in suboptimal clustering performance and indistinct facet mining results. To address the issue, we propose a novel paradigm shift from *Single-to-Multi* to *Multi-to-Multi* (**MoM**) clustering for multimedia data, which proactively integrates multiple informative views for decoupled multifaceted discovery, breaking the bottleneck of single-source data information scarcity. The proposed **MoM** framework consists of two core modules: the Dynamic Representation Fusion (DRF) module and the Decoupled Multifaceted Clustering (DMC) module. Specifically, the DRF module implements an adaptive fusion strategy that generates and dynamically fuses diverse representations guided by the anchors derived from a pretrained model, ensuring both fidelity and diversity. The DMC module leverages a structured self-expression model to jointly learn shared and facet-specific affinity matrices, producing well-separated and interpretable clusters. Unlike existing methods that passively decompose limited data into disjoint subspaces, our end-to-end framework actively employs a clustering-oriented representation fusion strategy to learn decoupled features tailored to each clustering facet. Extensive experiments show that **MoM** consistently outperforms state-of-the-art multifaceted clustering methods, validating the effectiveness of our **MoM** framework.

This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835249

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**.

Keywords

Deep Multifaceted Clustering, Dynamic Representation Fusion, Decoupled Representation Learning

ACM Reference Format:
Peipei Zhang, Wenjie Zhu, Qinli Zhou, and Weiqi Yan. 2026. MoM: Multi-to-Multi Identification of Decoupled Multifaceted Clustering with Dynamic Representation Fusion. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835249

1 Introduction

The explosive growth of multimedia data on platforms poses unprecedented challenges for content organization and understanding. Real-world multimedia instances often contain heterogeneous semantic cues, enabling multiple plausible grouping structures [3, 34]. For example, flower images can be clustered according to visual appearance such as color and texture, or semantics such as species or ecological attributes. Multifaceted clustering, also referred to as multiple or alternative clustering, seeks to uncover diverse and complementary partitions within a single dataset [49, 52]. This capability is pivotal for multimedia applications including content retrieval, recommendation, and media analytics, where users may require different organizational views based on their intent. Driven by the growing demand for flexible understanding of complex multimedia data beyond a single dominant clustering outcome, multifaceted clustering has recently attracted increasing attention as a promising paradigm for multimedia content analysis [40].

Multifaceted Clustering *vs* **Multi-view Clustering**. It is critical to distinguish multifaceted clustering from multi-view clustering, as the two paradigms differ fundamentally in input setting, methodology focus, and learning objective. Multi-view clustering [19, 33, 41, 54, 56] assumes that each instance is described by multiple heterogeneous features collected from different sources. Its core methodology lies in cross-view information fusion, aiming to

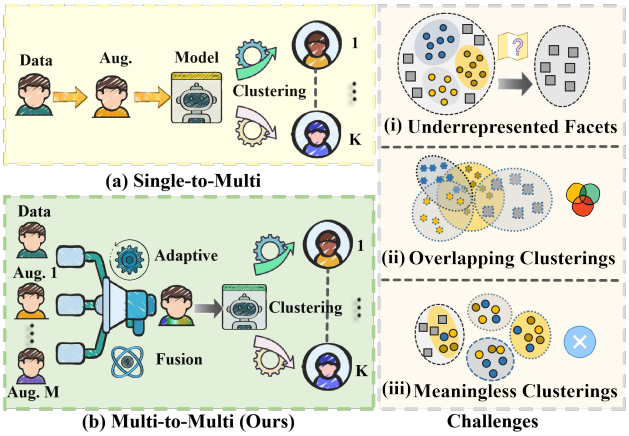


Figure 1: Illustration of challenges in multifaceted clustering.

integrate complementary information and derive a single consensus partition. In contrast, multifaceted clustering [7, 24, 32, 36, 38] operates in a single-view setting, where each instance is associated with only one representation. Instead of seeking consensus, its objective is to uncover diverse and interpretable clustering solutions, each capturing a distinct latent facet of the same data distribution. While multi-view clustering emphasizes information fusion across views, multifaceted clustering tackles the diversity extraction problem: deriving multiple non-redundant structural patterns from an inherently limited feature space.

Challenges. As shown in Figure 1, the challenge of multifaceted clustering lies not in fusing heterogeneous inputs, but in decoupling multiple latent heterogeneous structures within a unified yet limited feature space, whose representational capacity restricts the separation of multiple facets. Despite significant progress, existing multifaceted clustering methods still face three key challenges, which this work specifically addresses: **(i)** Conventional methods decompose the original single feature space into multiple non-redundant subspaces, yet remain constrained by intrinsic and dominant data patterns, failing to capture underrepresented facets. **(ii)** Sequential or unguided decomposition strategies may yield overlapping clusterings, due to the absence of explicit mechanisms ensuring facet distinctness. **(iii)** Augmentation-based methods often trade fidelity for diversity, leading to spurious clusters that are inconsistent with the real-world semantics underlying the data.

Novelty & Contributions. To address the above challenges, we propose a novel *Multi-to-Multi* (**MoM**) paradigm, which departs from the passive decomposition logic of previous methods and adopts an active *augmentation-fusion-discovery* pipeline. The core novelty lies in two aspects: **(i)** a dynamic representation fusion module that leverages anchor-guided augmentation to generate diverse yet meaningful views; and **(ii)** a decoupled multifaceted clustering module that jointly models shared and facet-specific affinities, reducing redundancy while preserving structural integrity. This paradigm shift enables **MoM** to actively expand the representational landscape beyond the intrinsic limits of the data itself, uncovering more diverse clustering facets. Different from multi-view clustering that fuses heterogeneous views for consensus learning, our fusion strategy aims to boost representation diversity and support structural decoupling. To sum up, the main contributions are summarized as follows:

- We propose a novel **MoM** clustering paradigm, which transforms single-view constrained multifaceted discovery into an active augmentation-fusion-discovery process, expanding the coverage of discoverable facets.
- We design a dynamic representation fusion module that employs anchor-guided augmentation to generate and adaptively fuse diverse clustering-oriented representations, achieving a balanced tradeoff between fidelity and diversity.
- We develop a decoupled multifaceted clustering module that learns shared and facet-specific affinity matrices via structured self-expression, effectively reducing redundancy and enhancing the distinctiveness of each discovered facet.
- Extensive experiments on multiple benchmarks demonstrate that **MoM** consistently surpasses state-of-the-art baselines in both clustering quality and diversity.

2 Related Work

2.1 Traditional Multifaceted Clustering

Traditional multifaceted clustering aims to uncover multiple valid yet diverse partitions from a single dataset, and can be broadly categorized into two main paradigms, namely simultaneous methods and sequential methods [13, 37, 48]. Simultaneous methods typically formulate multifaceted clustering as a multi-objective optimization problem that jointly balances clustering quality and facet diversity within a unified framework. To promote non-redundancy among the resulting partitions, these methods incorporate orthogonality or decorrelation constraints into classical clustering algorithms such as k-means or subspace clustering. Such constraints explicitly penalize similarity between alternative clusterings, often with respect to a reference partition [20, 22, 39, 59]. Sequential approaches [1], in contrast, generate alternative clusterings in an iterative manner. Each subsequent clustering is encouraged to differ from previous ones by either projecting data into subspaces that are uncorrelated with earlier solutions or by decomposing the data into independent latent factors via matrix factorization techniques [6, 43]. Despite their differences, a fundamental and persistent challenge shared across both paradigms lies in balancing partition quality, diversity, and interpretability. The evolution of the field demonstrates a shift from merely producing diverse clusterings to discovering complementary perspectives within the data.

2.2 Deep Multifaceted Clustering

Deep multifaceted clustering promotes traditional methods by leveraging deep neural networks to uncover multiple complementary partitions in nonlinear latent spaces. Core strategies encompass autoencoder-based representation learning with non-redundant objectives [25, 31, 44], as well as augmentation-based methods that generate diverse perspectives for improved clustering diversity [45, 53]. However, both categories remain inherently constrained by the intrinsic structure of the input data, limiting their capacity to reveal underrepresented clustering facets. Motivated by this limitation, vision-language model (VLM) [55] based approaches have been explored to incorporate semantic priors into multifaceted

clustering. By utilizing CLIP [27] embeddings and large language models [5, 17], these methods aim to align clustering structures with user-aligned semantics [35, 42, 46]. Despite their strong semantic expressiveness, such approaches are often limited by the single-to-multi paradigm, where multiple solutions are derived from a shared representation, restricting structural diversity. Furthermore, their performance typically depends on interactive feedback frameworks [15] or carefully designed prompts [16], limiting their generalizability. In contrast, our **MoM** framework emphasizes a data-centric approach to learn augmented representation fusion, enabling multifaceted clustering that addresses the inherent limitations of existing single-to-multi methods. Instead of passively decomposing a single representation, **MoM** actively constructs diverse yet meaningful views and jointly models their interactions, enabling the discovery of high quality and non-redundant clustering structures beyond the limitations of conventional single-to-multi paradigms.

3 Methodology

As outlined, our **MoM** comprises two core modules: **Dynamic Representation Fusion (DRF)** and **Decoupled Multifaceted Clustering (DMC)** modules. The pipeline is illustrated in Figure 2.

3.1 Preliminaries

Multifaceted clustering assumes that a dataset $\mathbf{X}$ can be partitioned into K distinct clustering solutions according to different criteria, where the K-th partition contains $|C_k|$ clusters. Formally, let the raw data be $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \in \mathbb{R}^{D\times N}$. To capture complementary information, we define an augmentation set $\mathcal{T} = \{\mathcal{T}_m\}_{m=1}^{M}$. Applying each $\mathcal{T}_m(\cdot)$ to $\mathbf{X}$ yields M augmented views $\mathcal{X} = \{\mathbf{X}^1, \cdots, \mathbf{X}^m, \cdots, \mathbf{X}^M\}$:

$$\mathbf{X}^m = \{\mathbf{x}_1^m, \mathbf{x}_2^m, \ldots, \mathbf{x}_N^m\} = \mathcal{T}_m(\mathbf{X}) \in \mathbb{R}^{D\times N}, \tag{1}$$

where $\mathbf{X}^m$ denotes the m-th augmented version of $\mathbf{X}$, and $\mathbf{x}_i^m$ is the corresponding transformed instance. Each augmentation induces a perturbed subspace of the data distribution, reflecting different invariance assumptions [2]. We adopt the image encoder of a pretrained model to obtain the corresponding representations:

$$\begin{aligned}\mathbf{F} &= \{\mathbf{f}_1, \cdots, \mathbf{f}_N\} = \{\mathcal{E}_{\text{img}}(\mathbf{x}_1), \cdots, \mathcal{E}_{\text{img}}(\mathbf{x}_N)\},\\ \mathbf{F}^m &= \{\mathbf{f}_1^m, \cdots, \mathbf{f}_N^m\} = \{\mathcal{E}_{\text{img}}(\mathbf{x}_1^m), \cdots, \mathcal{E}_{\text{img}}(\mathbf{x}_N^m)\},\end{aligned} \tag{2}$$

where $\mathcal{E}_{\text{img}}(\cdot)$ denotes the image encoder. The anchor associated to the j-th instance $\mathbf{x}_j$ is then computed as:

$$\mathbf{U}_j = \frac{1}{|\mathcal{S}_j|}\sum\nolimits_{\mathbf{x}_j\in\mathcal{S}_j} \frac{\mathbf{f}_j}{\|\mathbf{f}_j\|}. \tag{3}$$

Here, $\mathcal{S}_j$ denotes the index set of samples assigned to the same coarse semantic group as $\mathbf{x}_j$, which can be determined using the k-means algorithm to obtain coarse semantic groups. The anchor acts as the global semantic center, serving as a reference to align the consistency of augmented representation.

3.2 Dynamic Representation Fusion Module

Given the augmented representation in Eq. (2), the fused feature is denoted as $\widehat{\mathbf{F}} = \{\hat{\mathbf{f}}_1, \cdots, \hat{\mathbf{f}}_j, \cdots, \hat{\mathbf{f}}_N\} \in \mathbb{R}^{D\times N}$:

$$\hat{\mathbf{f}}_j = \sum\nolimits_{m=1}^{M} w_j^m \mathbf{f}_j^m, \tag{4}$$

where w_j^m denotes the weight assigned to the m-th augmentation of the j-th instance.

Parameterization of w_j^m in Eq. (4). We parameterize the weight w_j^m by explicitly integrating a global augmentation prior with instance-specific information. Specifically, we define a global distribution $\mathbf{p} = \{p^1, p^2, \ldots, p^M\} \in \mathbb{R}^M$, where p^m quantifies the overall effectiveness of the m-th augmentation across all instances. The global prior $\mathbf{p}$ is learnable and uniformly initialized, and is jointly updated through backpropagation with the fused representation objective. While $\mathbf{p}$ offers a macro-level assessment of augmentation utility, it overlooks instance-specific variations. To address this limitation, we introduce a sample-wise dynamic weight. The weight w_i^m in Eq. (4) is parameterized by combining the similarity s_i^m with the global prior p^m, as formulated below:

$$w_j^m = \frac{\exp\left(p^m \cdot s_j^m/\tau\right)}{\sum_{r=1}^{M}\exp\left(p^r \cdot s_j^r/\tau\right)}. \tag{5}$$

Here, $\tau > 0$ is a temperature parameter controlling the smoothness of the weight. For the weight matrix $\mathbf{W} \in \mathbb{R}^{N\times M}$, where each row $\mathbf{w}_j = \{w_j^1, \ldots, w_j^M\}$ is associated with the j-th instance, the dynamic element w_j^m reflects the contribution of the m-th augmentation to the j-th instance.

Parameterization of s_j^m in Eq. (5). The similarity between the m-th augmented representation and its corresponding anchor is computed as follows:

$$s_j^m = \text{sim}(\mathbf{f}_j^m, \mathbf{U}_j) = \frac{\mathbf{f}_j^m \cdot \mathbf{U}_j}{\|\mathbf{f}_j^m\| \cdot \|\mathbf{U}_j\|}, \tag{6}$$

where $\text{sim}(a, b)$ denotes cosine similarity, measuring how faithfully each augmented instance preserves the original semantic information. A higher score indicates strong semantic retention, while a lower score means feature distortion. Notably, the efficacy of a single augmentation varies across instances, demonstrating the instance-specific adaptability.

MoM Adapter. To fuse pretrained features while preserving the underlying subspace structure, we introduce a MoM adapter that maps the representation $\widehat{\mathbf{F}}$ into a latent space:

$$\widehat{\mathbf{Z}} = f(\widehat{\mathbf{F}}; \Theta_f), \quad s.t. \quad \|\widehat{\mathbf{Z}}\|_F^2 = N, \tag{7}$$

where $f(\cdot; \Theta_f)$ is a learnable three-layer MLP with ReLU that captures complementary representations across augmented views. Then an alignment objective regularizes it to preserve the semantic structure of pretrained features.

Representation Reconstruction. Rather than reconstructing pixel-level images, we adopt a representation-level reconstruction strategy to ensure semantic fidelity between the learned latent representations and the fused representations. Specifically, the reconstruction loss is defined as:

$$\mathcal{L}_{\text{RR}} = 1 - \frac{1}{N}\sum\nolimits_{j=1}^{N} \text{sim}\left(\widehat{\mathbf{z}}_j, \hat{\mathbf{f}}_j\right), \tag{8}$$

where $\widehat{\mathbf{z}}_j$ refers to the fused latent representation obtained by the MoM adapter in Eq. (7). This objective facilitates subspace information fusion and steers subspace learning.

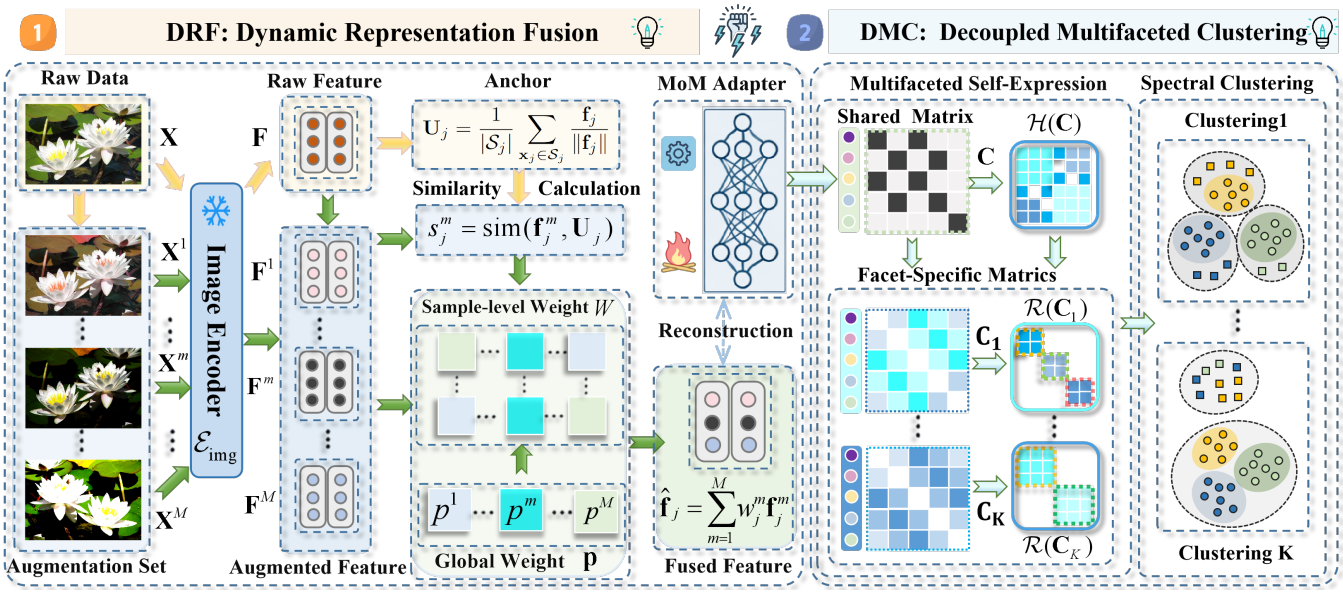


Figure 2: Overall pipeline of the MoM framework for multifaceted clustering, which integrates Dynamic Representation Fusion (DRF) and Decoupled Multifaceted Clustering (DMC) in an end-to-end manner. DRF enriches representations through anchor-guided adaptive fusion, while DMC exploits the fused features to jointly learn shared and facet-specific structures.

To avoid the trivial solution where a single augmentation dominates the fusion process, we introduce a maximum-entropy regularization term $\mathcal{H}(\mathbf{W}) \triangleq -\sum_{j=1}^{N}\sum_{m=1}^{M} w_j^m \log w_j^m$. Maximizing the entropy of the weight matrix $\mathbf{W}$ encourages balanced contributions from all augmentations instead of extreme weighting. The overall objective, weighted by the hyperparameter η, is defined as:

$$\mathcal{L}_{\text{DRF}} = \mathcal{L}_{\text{RR}} - \eta \cdot \mathcal{H}(\mathbf{W}). \tag{9}$$

Due to the additivity of $\mathcal{L}_{\text{RR}}$ and $\mathcal{H}(\mathbf{W})$ across samples, minimizing $\mathcal{L}_{\text{DRF}}$ reduces to optimizing the instance-level regularized objective for each sample j. The optimal w_j^m is given by:

Theorem 1 (Optimal Dynamic Weighting). *Let $\mathcal{J}(w_j)$ denote the instance-level objective defined in Eq. (9). The optimal weight w_j^m that minimizes $\mathcal{J}(w_j)$ admits a closed-form softmax solution proportional to* $\exp(\langle \mathbf{f}_j^m, \mathbf{U}_j \rangle / \tau)$.

Theorem 1 shows that the optimal w_j^m follow a softmax form distribution over anchor similarity, providing a theoretical justification for the adaptive fusion mechanism in $\mathcal{L}_{\text{DRF}}$.

3.3 Decoupled Multifaceted Clustering Module

We assume that multifaceted clustering can be represented by shared and facet-specific self-expression matrices. The shared matrix captures global cluster structures, while the facet-specific matrices retain the unique characteristics of each subspace. Our goal is to jointly model both components to recover consistent global partitions while preserving discriminative information.

Shared Self-Expression. Let $\mathbf{C} = \{c_{i,j}\}_{i,j=1}^{N} \in \mathbb{R}^{N\times N}$ denote the shared self-expression matrix, where each element $c_{i,j}$ quantifies the contribution of instance j to reconstructing instance i. The loss for learning $\mathbf{C}$ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{SSE}} &= \alpha \|\widehat{\mathbf{Z}} - \widehat{\mathbf{Z}}\mathbf{C}\|_F^2 + \beta \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}\log c_{i,j}, \\ s.t. \quad & \operatorname{diag}(\mathbf{C}) = 0, \quad c_{i,j} \geq 0, \end{aligned} \tag{10}$$

where the hyperparameters α and β are introduced to balance the relative importance between the reconstruction term and the regularization term. Meanwhile, a zero-diagonal is enforced on the matrix $\mathbf{C}$ to avoid the trivial identity mapping [57, 58]. To further promote diverse affinity patterns, we impose maximum entropy regularization, which avoids overly sparse or deterministic affinities and thus enhances the diversity of relational information.

Facet-specific Self-Expression. While shared matrix $\mathbf{C}$ models global pairwise relations, it may entangle distinct latent structures. To decouple them, we introduce a facet-specific self-expression learning scheme that learns a set of matrices $\{\mathbf{C}_k\}_{k=1}^{K}$, each corresponding to a distinct clustering facet. Parallel self-expression layers with block-diagonal regularization facilitate structural diversity and decoupling:

$$\begin{aligned} \mathcal{L}_{\text{FSE}} &= \sum_{k=1}^{K} \alpha \left(\|\mathbf{C} - \mathbf{C}_k\|_F^2 + \|\widehat{\mathbf{Z}} - \widehat{\mathbf{Z}}\mathbf{C}_k\|_F^2 \right) + \beta \mathcal{R}(\mathbf{C}_k), \\ s.t. \quad & \operatorname{diag}(\mathbf{C}_k) = \mathbf{0}, k \in \{1, \ldots, K\}, \end{aligned} \tag{11}$$

where $\mathcal{R}(\cdot)$ denotes the regularization term, and matrices $\{\mathbf{C}_k\}_{k=1}^{K}$ are produced by K independent fully connected layers. To reinforce the structural diversity among different matrices $\{\mathbf{C}_k\}_{k=1}^{K}$, we adopt the block-diagonal regularization [18, 23]. Giving the facet-specific affinity matrices $\{\mathbf{A}_k\}_{k=1}^{K} = \{(|\mathbf{C}_k| + |\mathbf{C}_k|^\top)/2\}_{k=1}^{K}$,

Algorithm 1 Procedure of the proposed **MoM** framework.

Require: Data matrix $\mathbf{X}$, augmentation set $\mathcal{T} = \{\mathcal{T}_m\}_{m=1}^{M}$, tradeoff parameters α, β, η, τ, the number of epochs T, clustering solutions K, and clusters for each solution $|C_k|$.

Ensure: Final cluster assignments for all data samples.

1: Obtain the pretrained feature representation via Eq. (2) and the anchor via Eq. (3).
2: Calculate the anchor similarity via Eq. (6).
3: Initialize the shared self-expression matrix $\mathbf{C}$ and the set of facet-specific self-expression matrices $\{\mathbf{C}_k\}_{k=1}^{K}$.
4: **for** each epoch = 1 to T **do**
5: Update the network parameters via Eq. (13).
6: **end for**
7: Generate the set of multiple affinity matrices $\{\mathbf{A}_k\}_{k=1}^{K}$ from the learned self-expression matrices.
8: **Clustering Stage:** Perform spectral clustering on $\{\mathbf{A}_k\}_{k=1}^{K}$ to obtain the final assignments.

the regularization is expressed as:

$$\mathcal{R}(\mathbf{C}_k) = \|\mathbf{A}_k\|_{|C_k|} = \sum_{j=N-|C_k|+1}^{N} \lambda_j(\mathbf{L}_k), \tag{12}$$

where $\mathbf{L}_k = \mathbf{D}_k - \mathbf{A}_k \in \mathbb{R}^{N\times N}$ is the Laplacian with degree matrix $\mathbf{D}_k \in \mathbb{R}^{N\times N}$, and λ_j is the j-th largest eigenvalue. **MoM** does not enforce complete independence among facets, as meaningful facets naturally share partial semantic cues. By imposing facet-specific block-diagonal structures aligned with $|C_k|$, the model learns heterogeneous and complementary matrices across facets.

The overall training loss combines the dynamic representation fusion and the decoupled multifaceted clustering module:

$$\mathcal{L} = \mathcal{L}_{\text{DRF}} + \mathcal{L}_{\text{SSE}} + \mathcal{L}_{\text{FSE}}. \tag{13}$$

Once the facet-specific matrices $\{\mathbf{C}_k\}_{k=1}^{K}$ are obtained, spectral clustering is performed on the affinity matrices $\{\mathbf{A}_k\}_{k=1}^{K}$[4, 26, 30, 51], yielding diverse and non-redundant clusterings. For clarity, the complete training procedure is summarized in Algorithm 1.

4 Experiments

In this section, we conduct a comprehensive evaluation of our proposed **MoM** by addressing the following key questions:

- **Q1: Performance Evaluation**. Does **MoM** outperform state-of-the-art methods in multifaceted clustering tasks?
- **Q2: Ablation Study**. Which of the individual loss components play a critical role in boosting the overall performance of **MoM**?
- **Q3: Backbone Impact**. What is the impact of different pretrained models on the final clustering quality?
- **Q4: Fusion**. How do fused features differ from raw image features in multifaceted clustering tasks?
- **Q5: Hyperparameter Sensitivity**. To what extent do hyperparameter variations affect the stability of **MoM**?
- **Q6: Interpretability Assessment**. Are the discovered facets interpretable and semantically meaningful?

4.1 Experiment Setting

Datasets. We evaluate our proposed **MoM** on six benchmark datasets, including ALOI [8], Fruit [12], Fruit360, Cards [45], Flowers and Stanford Cars [47] datasets.

Baselines. To ensure a comprehensive comparison of performance, we evaluate our **MoM** framework against a suite of multifaceted clustering baselines, including traditional multifaceted clustering approaches OSC [6], Nr-Kmeans [21], and MSC [12], and deep learning–based multifaceted clustering methods MCV [9], ENRC [25], iMClusts [31], AugDMC [45], and CMClusts [53].

Metrics. Clustering performance is evaluated using two metrics: Normalized Mutual Information (NMI) and Rand index (RI) [10, 50].

Implementation Details. We employ several augmentations spanning color, geometric, edge, gamma, texture, noise, and contrast transformations to promote appearance diversity, spatial invariance and other desirable properties. The same strategy is applied across all datasets. For optimization, we adopt the Adam optimizer with ReLU activation. The hyperparameters are set as follows: $\eta = 1e^3$ and $\tau = 0.1$ for all datasets. The learning rate is selected from $\{1e^{-4}, 5e^{-3}, 1e^{-3}\}$, and the weight decay is tuned within $\{5e^{-3}, 1e^{-3}, 5e^{-4}\}$. The trade-off parameters α and β are searched over $\{1e^{-3}, 1e^{-2}, 1e^{-1}, 1, 1e^{1}, 1e^{2}, 1e^{3}\}$. We adopt CLIP ViT-B/32 [14, 28] as the pretrained image encoder for all datasets. All experiments are repeated ten times with different random seeds, and the reported results are averaged over these runs.

4.2 Clustering Performance Analysis (Q1)

We evaluate **MoM** against a suite of state-of-the-art baselines across six datasets. Across all evaluation metrics, **MoM** consistently achieves the best performance, suggesting that the augmentation–fusion–discovery mechanism effectively expands the representational capacity for uncovering diverse and meaningful facets. Table 1 reports the detailed NMI and RI results across all datasets and methods, from which three key observations can be drawn:

(i) Superiority of Deep Architectures. Shallow methods often struggle to model the nonlinear relationships inherent in data, whereas deep architectures learn discriminative representations that align with the intrinsic structure of the data. Consistently, deep models outperform all shallow methods across all evaluation metrics. Specifically, on the ALOI dataset, all deep models achieve nearly perfect clustering scores, which fully demonstrates the strong representational capacity of deep architectures.

(ii) Benefit of End-to-End Optimization. Two-stage or EM-based methods often yield suboptimal features, as the processes of feature extraction and clustering are separated. In contrast, our end-to-end framework jointly optimizes representation learning and clustering in a unified manner. It leverages the designed multifaceted clustering loss to guide dynamic representation fusion process, ensuring that features are tailored to multifaceted clustering tasks and supporting the discovery of diverse facets.

(iii) Effectiveness of the MoM Paradigm. While existing augmentation-based methods improve clustering performance, their fusion strategies remain static and limited. Our adaptive **MoM** paradigm dynamically fuses diverse augmented views, uncovering complementary facets beyond the data's intrinsic structure. Overall, **MoM** achieves the best performance, outperforming CMClusts,

Table 1: Clustering results in terms of NMI and RI on six datasets. The best and second-best results are highlighted with darker and lighter gray shades, respectively.

Dataset		ALOI		Fruit		Fruit360		Cards		Flowers		Stanford Cars	
Method	Metric	Shape	Color	Color	Species	Color	Species	Number	Suits	Color	Species	Color	Type
OSC	NMI↑	0.3106	0.2257	0.3763	0.1104	0.3053	0.3105	0.1095	0.0359	0.1876	0.1370	0.2366	0.1047
	RI↑	0.6112	0.5834	0.6711	0.5599	0.6716	0.6726	0.7820	0.3696	0.4821	0.4506	0.6141	0.5072
Nr-Kmeans	NMI↑	0.2831	0.2312	0.4665	0.1351	0.3768	0.3839	0.1126	0.0270	0.2166	0.1514	0.2451	0.1191
	RI↑	0.6046	0.5793	0.7317	0.5679	0.6899	0.6958	0.7651	0.3564	0.4848	0.5482	0.6377	0.5151
MSC	NMI↑	0.2968	0.1563	0.6886	0.1627	0.2544	0.2184	0.0807	0.0497	0.2561	0.1326	0.2331	0.1325
	RI↑	0.5199	0.3428	0.8051	0.6045	0.6054	0.5805	0.7805	0.3587	0.5965	0.5273	0.6158	0.5336
MCV	NMI↑	0.7359	0.6982	0.6266	0.2733	0.3776	0.2985	0.0792	0.0430	0.2938	0.1561	0.2103	0.1650
	RI↑	0.8261	0.7439	0.7685	0.6597	0.6791	0.6176	0.7128	0.3638	0.5860	0.6065	0.5802	0.5634
ENRC	NMI↑	0.9732	0.9833	0.7103	0.3197	0.4264	0.4142	0.1225	0.0676	0.3329	0.1894	0.2465	0.2063
	RI↑	0.9861	0.9892	0.8511	0.6536	0.6868	0.6984	0.7313	0.3801	0.6214	0.6195	0.6779	0.6217
iMClusts	NMI↑	0.9963	**1.0000**	0.7351	0.3029	0.4097	0.3861	0.1144	0.0716	0.3169	0.1887	0.2336	0.1963
	RI↑	0.9989	**1.0000**	0.8632	0.6743	0.6841	0.6732	0.7658	0.3715	0.6127	0.6077	0.6552	0.5643
AugDMC	NMI↑	**1.0000**	**1.0000**	0.8517	0.3546	0.4594	0.5139	0.1440	0.0873	0.3556	0.1996	0.2736	0.2364
	RI↑	**1.0000**	**1.0000**	0.9108	0.7399	0.7392	0.7430	0.8267	0.4228	0.6931	0.6227	0.7525	0.7356
CMClusts	NMI↑	**1.0000**	**1.0000**	0.7130	0.2170	0.5732	0.4087	0.2263	0.1782	0.3654	0.1535	0.2464	0.2047
	RI↑	**1.0000**	**1.0000**	0.8615	0.6143	0.7508	0.7252	0.6078	0.4691	0.7136	0.5714	0.7363	0.7161
MoM (Ours)	NMI↑	**1.0000**	**1.0000**	**0.9611**	**0.8920**	**0.5746**	**0.5715**	**0.2512**	**0.3110**	**0.4663**	**0.4265**	**0.4673**	**0.5141**
	RI↑	**1.0000**	**1.0000**	**0.9892**	**0.9621**	**0.7562**	**0.7532**	**0.8718**	**0.7017**	**0.7748**	**0.7462**	**0.7689**	**0.8243**

AugDMC, and all other competitors, confirming the superiority of its unified augmentation–fusion–discovery framework.

4.3 Ablation Studies (Q2, Q3, Q4)

Influence of two key modules (Answer Q2). We perform ablation studies on the Fruit360, Cards, and Stanford Cars datasets using CLIP features to assess the contribution of each loss component. Specifically, Variant V1 applies direct CLIP + k-means clustering; V2 employs fused features + k-means; V3 denotes a standard deep subspace clustering adopting a shared matrix across all clusterings; and **MoM** denotes the full model with both $\mathcal{L}_{\text{SSE}}$ and $\mathcal{L}_{\text{FSE}}$. As shown in Table 2, the results reveal that k-means fails to uncover multifaceted patterns, whereas self-expression learning decouples orthogonal subspaces for improved facet separation. Moreover, excluding $\mathcal{L}_{\text{FSE}}$ leads to diminished diversity and consistent performance drops across the evaluated datasets, underscoring that the diverse facets captured by **MoM** arise from the synergistic interaction between representation fusion and decoupled multifaceted clustering.

Impact of pretrained model choice (Answer Q3). We evaluate the impact of pretrained backbones on **MoM** using ResNet18, ResNet34, and ResNet50 [11], together with CLIP variants (ViT-L/14, ViT-B/16, and ViT-B/32) [29]. As illustrated in Figure 3 on the Fruit dataset, pretrained encoders with stronger representation capabilities generally yield higher clustering quality and more stable facet diversity. Notably, even when equipped with lightweight backbones, **MoM** still maintains competitive performance, which demonstrates its adaptability to different pretrained initializations.

Table 2: Ablation studies in terms of NMI and RI. The best and second-best results are highlighted with darker and lighter gray shades, respectively.

Method	Loss Terms			Datasets	Fruit360		Cards		Stanford Cars	
Variants	$\mathcal{L}_{\text{DRF}}$	$\mathcal{L}_{\text{SSE}}$	$\mathcal{L}_{\text{FSE}}$	Metric	Color	Species	Number	Suits	Color	Type
V1	×	×	×	NMI↑	0.4060	0.3799	0.1098	0.0963	0.0582	0.1825
				RI↑	0.7099	0.6875	0.8186	0.6836	0.6370	0.6099
V2	✓	×	×	NMI↑	0.4597	0.4355	0.1406	0.1184	0.0438	0.2764
				RI↑	0.7270	0.7160	0.8261	0.6854	0.6166	0.6661
V3	✓	✓	×	NMI↑	0.5591	0.4203	0.2427	0.1252	0.3585	0.2205
				RI↑	0.7550	0.7053	0.8616	0.6505	0.7413	0.6130
MoM	✓	✓	✓	NMI↑	**0.5746**	**0.5715**	**0.2512**	**0.3110**	**0.4673**	**0.5141**
				RI↑	**0.7562**	**0.7532**	**0.8718**	**0.7017**	**0.7689**	**0.8243**

Importance of fused representation (Answer Q4). As shown in Figure 4, we compare our fused representation with CLIP+k-means, Adapter+k-means, and our proposed **MoM**, each evaluated with and without the proposed fusion strategy on the Fruit360 dataset. Clustering based solely on the original feature remains confined to surface similarities, leading to overlapping groups. In contrast, our fusion method transcends limitations by dynamically integrating complementary information from multiple augmented

views, yielding well-structured clusters. This demonstrates that **MoM** expands the representational landscape beyond the data itself, uncovering richer and more diverse facets of underlying structure.

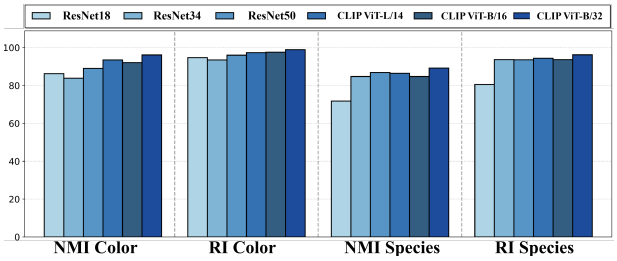


Figure 3: Pretrained model comparison on the Fruit dataset.

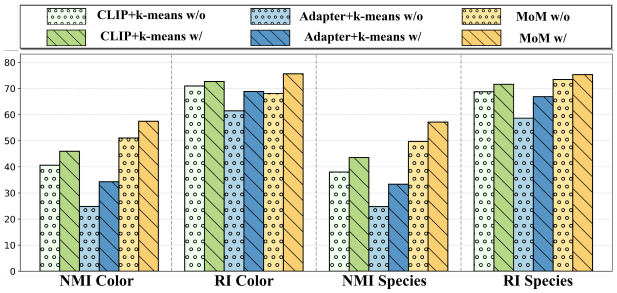


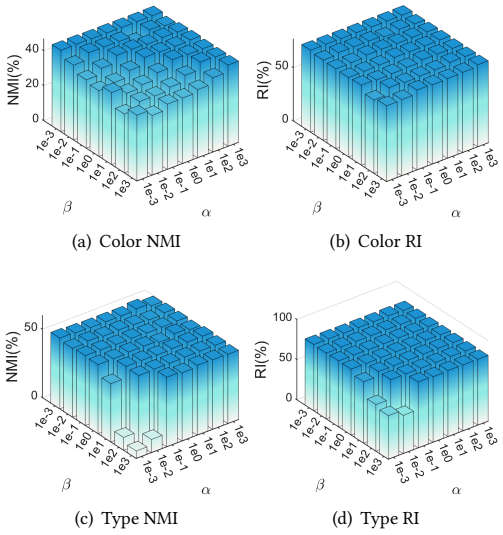
Figure 4: Fruit360 results with and w/o fused features.

4.4 Parameter Sensitivity Analysis (Q5)

To further investigate the impact of hyperparameters on clustering performance, we conduct a sensitivity analysis on the Stanford Cars dataset, with $\alpha, \beta \in \{1e^{-3}, 1e^{-2}, 1e^{-1}, 1, 1e^{1}, 1e^{2}, 1e^{3}\}$. As illustrated in Figure 5, **MoM** demonstrates consistently stable performance across a broad range of parameter configurations, indicating low sensitivity to hyperparameter choices. Performance degrades only when α is set too small, diminishing the impact of the self-expression loss or when β is overly large, leading to excessive regularization. The optimal performance is obtained when α and β are set to moderate values, balancing reconstruction, self-expression, and regularization. Importantly, we observe that **MoM** achieves nearly optimal performance when both α and β are set within a moderate range, where a favorable balance between reconstruction, self-expression, and regularization is maintained. Based on these observations, we fix α and β within this range for all experiments.

4.5 Clustering Interpretability Evaluation (Q6)

Cluster Centroid Inspection. We assess the clustering interpretability by visualizing cluster centroids on the Fruit dataset. For multifaceted clustering, interpretability demands that each clustering solution yields a distinct and meaningful partition. Specifically, we construct cluster centroids by aggregating samples with identical cluster labels across all clusterings. As shown in Figure 6, panels (a)–(c) present color-based clusters with distinct red, green, and yellow hues, while panels (d)–(f) show species-based clusters



(a) Color NMI (b) Color RI

(c) Type NMI (d) Type RI

Figure 5: Clustering results in terms of NMI and RI versus different choices of α and β on the Stanford Cars dataset.

that exhibit minimal color overlap. These findings substantiate that **MoM** effectively disentangles diverse clustering-oriented facets, resulting in interpretable and distinguishable cluster structures.



(a) Red (b) Green (c) Yellow

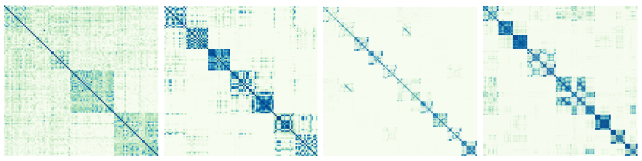


(d) Apple (e) Banana (f) Grape

Figure 6: Visualization of cluster centroids on Fruit dataset.

Visualization of Affinity Matrices. To further validate the self-expression capacity of **MoM**, we visualize the learned facet-specific affinity matrices across multiple benchmark datasets. As observed in Figure 7, the resulting matrices exhibit two desirable properties that are essential for self-expression learning-based multifaceted clustering: **(i) Block-diagonal structure**: The matrices exhibit clear block-diagonal patterns, which reflect the inherent subspace-preserving characteristic. This demonstrates that **MoM** can strengthen the connections among samples within the same subspace while alleviating interferences across different subspaces.

(ii) Diverse and interpretable patterns: Each matrix highlights distinct and semantically coherent clustering behaviors aligned with underlying grouping criteria. For instance, on the Fruit and Fruit360 datasets, **MoM** distinctly separates color-based clusters from species-based clusters, revealing its capability to unveil heterogeneous multifaceted structures embedded within complex data.



(a) Color (b) Species (c) Color (d) Species

Figure 7: Visualization of the affinity matrices. Panels (a)–(b) show the Fruit dataset, and (c)–(d) the Fruit360 dataset.

Analysis of weight p. Figure 8 shows the adaptive fusion weights **p** on the Cards dataset. In panel (a), intra-class patterns such as character decorations create interference. Edge Detection focuses on the essential contours of suits and numbers, suppressing these distractions. Gaussian Blur smooths minor textures yet preserves core structures, enhancing feature stability. Geometric Augmentation mitigates spatial variations and Color Jitter compensates for color and lighting changes, both making moderate contributions. Basic and Contrast Enhancement provide little useful variation or distort key features, thus receive low weights. Gaussian Noise rises in later epochs, introducing controlled perturbations that regularize the model once core features are established. The learned weights strike a balance between emphasizing discriminative structures, preserving feature stability, and suppressing irrelevant interference.

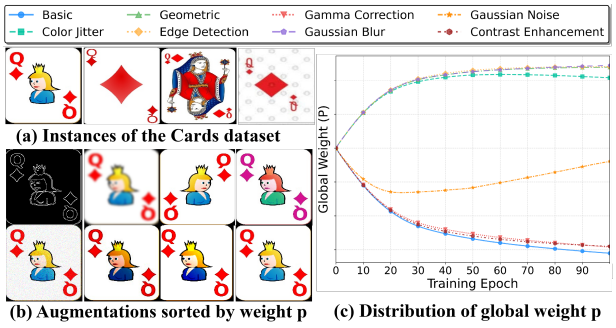


Figure 8: Priority ranking of global weight p on the Cards.

Interpretability of weight w_j^m. We further analyze the weight w_j^m shown in Figure 9, which reflects the model's dynamic fusion strategy guided by the anchor. When attribute conflicts arise, such as identical species with different colors, the model suppresses color-based perturbations while emphasizing contour and texture cues, thereby remaining invariant to superficial chromatic variations. Conversely, when color and species attributes are consistent, the model increases the contribution of color-related transformations to preserve discriminative information under geometric changes. Such sample-level adaptability enables **MoM** to achieve interpretable and clustering-oriented augmentation selection, effectively decoupling intrinsic cluster structures from multifaceted clustering.

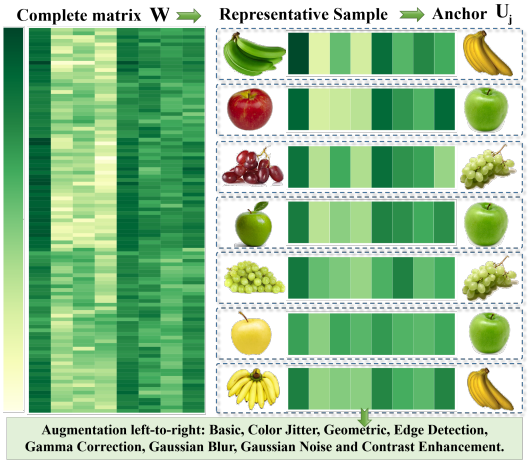


Figure 9: Interpretability of weights w_j^m on the Fruit dataset.

4.6 Efficiency and scalability.

As shown in Table 3, we compare the efficiency of **MoM** with representative multifaceted clustering methods on Fruit360. **MoM** learns one shared and K facet-specific $N \times N$ matrices, resulting in a dominant complexity of $O((K+1)N^2)$. It increases memory usage yet reflects a memory-time trade-off: **MoM** trains one shared encoder for all facets, while AugDMC and CMClusts require multiple encoders or repeated training. Thus, **MoM** is faster in practice.

Table 3: Efficiency and complexity analysis on Fruit360.

Methods	Parameters	Running Time	Peak Memory
AugDMC	13.27M	4848.92 s	903.96 MB
CMClusts	48.14M	352.02 s	4501.43 MB
MoM (Ours)	72.84M	**131.03 s**	6871.20 MB

5 Conclusion

In this paper, we introduced **MoM**, a novel Multi-to-Multi multifaceted clustering framework that actively integrates multiple informative views for decoupled facet discovery. By combining dynamic representation fusion to generate diverse and semantically aligned representations with decoupled multifaceted clustering for learning shared and facet-specific affinities, our method produces well-separated, interpretable clusters. Extensive evaluations demonstrate that **MoM** consistently outperforms existing multifaceted clustering approaches, confirming the advantages of active representation fusion over passive subspace decomposition.

Despite its effectiveness in single-modal multifaceted clustering, **MoM** is limited to unimodal data due to its dynamic representation fusion strategy. Future work will extend **MoM** to multimodal multimedia scenarios, enabling the exploitation of complementary cross-modal information for more comprehensive facet discovery.

References

[1] Eric Bae and James Bailey. 2006. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *Sixth International Conference on Data Mining*. 53–62.

[2] Chunchun Chen, Xing Wei, Jiayi Yang, Chenrun Wang, Yiwei Fu, Yuxing Zhang, Xin Sun, Rui Fan, and Wei Ye. 2026. Edge Self-Adversarial Augmentation Enhances Graph Contrastive Learning Against Neighborhood Inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 20005–20013.

[3] Jie Chen, Hua Mao, Wai Lok Woo, Chuanbin Liu, Zhu Wang, and Xi Peng. 2025. One-Step Adaptive Graph Learning for Incomplete Multiview Subspace Clustering. *IEEE Transactions on Knowledge and Data Engineering* 37, 5 (2025), 2771–2783.

[4] Siyu Chen, Xiaoqian Zhang, Youdong He, Lifan Peng, Yanchi Ou, and Shijie Xu. 2025. Multi-dimensional weighted deep subspace clustering with feature classification. *Expert Systems with Applications* 260 (2025), 125375.

[5] Ziye Chen, Yiqun Duan, Riheng Zhu, Zhenbang Sun, and Mingming Gong. 2025. Agent-Centric Personalized Multiple Clustering with Multi-Modal LLMs. *arXiv preprint arXiv:2503.22241* (2025).

[6] Ying Cui, Xiaoli Z Fern, and Jennifer G Dy. 2007. Non-redundant multi-view clustering via orthogonalization. In *Seventh IEEE International Conference on Data Mining*. IEEE, 133–142.

[7] Xiwen Geng, Suyun Zhao, Yixin Yu, Borui Peng, Pan Du, Hong Chen, Cuiping Li, and Mengdie Wang. 2025. Personalized clustering via targeted representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 16790–16798.

[8] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. 2005. The Amsterdam library of object images. *International Journal of Computer Vision* 61, 1 (2005), 103–112.

[9] Joris Guérin and Byron Boots. 2018. Improving Image Clustering With Multiple Pretrained CNN Feature Extractors. In *British Machine Vision Conference, BMVC*.

[10] Wengang Guo, Wei Ye, Chunchun Chen, Xin Sun, Christian Böhm, Claudia Plant, and Susanto Rahardja. 2025. Bootstrap Deep Spectral Clustering with Optimal Transport. *IEEE Transactions on Multimedia* 28 (2025), 531–544.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[12] Juhua Hu, Qi Qian, Jian Pei, Rong Jin, and Shenghuo Zhu. 2017. Finding multiple stable clusterings. *Knowledge and Information Systems* 51, 3 (2017), 991–1021.

[13] Collin Leiber, Lukas Miklautz, Claudia Plant, and Christian Böhm. 2023. Benchmarking deep clustering algorithms with clustpy. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 625–632.

[14] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. 2024. Image clustering with external guidance. In *Proceedings of the 41st International Conference on Machine Learning*. 27890–27902.

[15] Honglin Liu, Peng Hu, Changqing Zhang, Yunfan Li, and Xi Peng. 2024. Interactive deep clustering via value mining. *Advances in Neural Information Processing Systems* 37 (2024), 42369–42387.

[16] Honglin Liu, Chao Sun, Peng Hu, Yunfan Li, and Xi Peng. 2026. Conditional Representation Learning for Customized Tasks. *Advances in Neural Information Processing Systems* 38 (2026), 31706–31737.

[17] Mingxuan Liu, Zhun Zhong, Jun Li, Gianni Franchi, Subhankar Roy, and Elisa Ricci. 2026. Organizing unstructured image collections using natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8907–8918.

[18] Canyi Lu, Jiashi Feng, Zhouchen Lin, Tao Mei, and Shuicheng Yan. 2018. Subspace clustering by block diagonal representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 487–501.

[19] Caixuan Luo, Jie Xu, Yazhou Ren, Junbo Ma, and Xiaofeng Zhu. 2024. Simple contrastive multi-view clustering with data-level fusion. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 4697–4705.

[20] Shirui Luo and Xiaochun Cao. 2021. Multiview subspace dual clustering. *IEEE Transactions on Neural Networks and Learning Systems* 33, 12 (2021), 7425–7437.

[21] Dominik Mautz, Wei Ye, Claudia Plant, and Christian Böhm. 2018. Discovering non-redundant k-means clusterings in optimal subspaces. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1973–1982.

[22] Dominik Mautz, Wei Ye, Claudia Plant, and Christian Böhm. 2020. Non-redundant subspace clusterings with nr-kmeans and nr-dipmeans. *ACM Transactions on Knowledge Discovery from Data* 14, 5 (2020), 1–24.

[23] Xianghan Meng, Zhiyuan Huang, Wei He, Xianbiao Qi, Rong Xiao, and Chun-Guang Li. 2025. Exploring a Principled Framework for Deep Subspace Clustering. In *International Conference on Learning Representations*, Vol. 2025. 78102–78135.

[24] Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, and Ioannis Patras. 2023. Divclust: Controlling diversity in deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3418–3428.

[25] Lukas Miklautz, Dominik Mautz, Muzaffer Can Altinigneli, Christian Böhm, and Claudia Plant. 2020. Deep embedded non-redundant clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5174–5181.

[26] Bo Peng, Yuanwei Hu, Bo Liu, Ling Chen, Jie Lu, and Zhen Fang. 2026. Delving into Spectral Clustering with Vision-Language Representations. In *The Fourteenth International Conference on Learning Representations*.

[27] Bo Peng, Jie Lu, Zhen Fang, and Guangquan Zhang. 2026. A Close Look at Negative Label Guided Out-of-distribution Detection in Pre-trained Vision-Language Models. In *Proceedings of the Forty-third International Conference on Machine Learning*.

[28] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. 2025. On the Provable Importance of Gradients for Autonomous Language-Assisted Image Clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19805–19815.

[29] Bo Peng, Jie Lu, Guangquan Zhang, and Zhen Fang. 2026. Debiased negative mining improves out-of-distribution detection with pre-trained vision-language models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1171–1181.

[30] Bo Peng and Wenjie Zhu. 2021. Deep structural contrastive subspace clustering. In *Asian Conference on Machine Learning*. PMLR, 1145–1160.

[31] Liangrui Ren, Guoxian Yu, Jun Wang, Lei Liu, Carlotta Domeniconi, and Xiangliang Zhang. 2023. A diversified attention model for interpretable multiple clusterings. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2023), 8852–8864.

[32] Jun Wang, Huiling Zhang, Wei Ren, Maozu Guo, and Guoxian Yu. 2021. EpiMC: detecting epistatic interactions using multiple clusterings. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 1 (2021), 243–254.

[33] Qianqian Wang. 2026. LLM-DAMVC: A Large Language Model Assisted Dynamic Agent for Multi-View Clustering. *Advances in Neural Information Processing Systems* 38 (2026), 119515–119534.

[34] Qianqian Wang, Haiming Xu, Wei Feng, and Quanxue Gao. 2026. Adversarial Fair Incomplete Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 10011–10019.

[35] Xinyue Wang, Yuheng Jia, Hui Liu, and Junhui Hou. 2026. Esmc: Mllm-based embedding selection for explainable multiple clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 26588–26596.

[36] Shaowei Wei, Jun Wang, Guoxian Yu, Carlotta Domeniconi, and Xiangliang Zhang. 2020. Deep incomplete multi-view multiple clusterings. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 651–660.

[37] Shaowei Wei, Jun Wang, Guoxian Yu, Carlotta Domeniconi, and Xiangliang Zhang. 2020. Multi-view multiple clusterings using deep matrix factorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6348–6355.

[38] Shaowei Wei, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. 2021. Multiple clusterings of heterogeneous information networks. *Machine Learning* 110, 6 (2021), 1505–1526.

[39] Xiaolong Xiong, Jinhan Cui, Jiaxiong Liu, Shuzhan Guo, and Jun Zhou. 2024. Inverse optimization for multi-view multiple clustering. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[40] Xiaolong Xiong, Jinhan Cui, Rui Xie, Shuzhan Guo, and Jun Zhou. 2024. Large-scale multi-view multiple clustering. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6205–6209.

[41] HaiMing Xu, Qianqian Wang, Boyue Wang, and Quanxue Gao. 2025. Deep fair multi-view clustering with attention KAN. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5061–5070.

[42] Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Ziyue Peng, Zewei Liu, Hewei Wang, Jiayi Zhang, and Edith CH Ngai. 2026. Multi-modal Dynamic Proxy Learning for Personalized Multiple Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 27269–27277.

[43] Sen Yang and Lijun Zhang. 2017. Non-redundant multiple clustering by nonnegative matrix factorization. *Machine Learning* 106, 5 (2017), 695–712.

[44] Jiawei Yao and Juhua Hu. 2024. Dual-disentangled deep multiple clustering. In *Proceedings of the 2024 SIAM International Conference on Data Mining*. SIAM, 679–687.

[45] Jiawei Yao, Enbei Liu, Maham Rashid, and Juhua Hu. 2023. Augdmc: Data augmentation guided deep multiple clustering. *Procedia Computer Science* 222 (2023), 571–580.

[46] Jiawei Yao, Qi Qian, and Juhua Hu. 2024. Customized multiple clustering via multi-modal subspace proxy learning. *Advances in Neural Information Processing Systems* 37 (2024), 82705–82725.

[47] Jiawei Yao, Qi Qian, and Juhua Hu. 2024. Multi-modal proxy learning towards personalized visual multiple clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14066–14075.

[48] Shixin Yao, Guoxian Yu, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. 2019. Multi-view multiple clustering. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 4121–4127.

[49] Guoxian Yu, Liangrui Ren, Jun Wang, Carlotta Domeniconi, and Xiangliang Zhang. 2024. Multiple clusterings: Recent advances and perspectives. *Computer Science Review* 52 (2024), 100621.

[50] Junpu Zhang, Shengju Yu, Suyuan Liu, Siwei Wang, Miaomiao Li, Xinwang Liu, En Zhu, and Kunlun He. 2025. Learning the Anchors with Similar Distributions to Original Data for Multi-view Clustering. In *Proceedings of the 33rd ACM*

International Conference on Multimedia. 857–866.

[51] Peipei Zhang, Wenjie Zhu, and Wei Qi Yan. 2025. Multi-Level Structural Contrastive Subspace Clustering Network. *IEEE Signal Processing Letters* 32 (2025), 3092–3096.

[52] Shan Zhang, Liangrui Ren, Qiaoyu Tan, Carlotta Domeniconi, Wei Du, Jun Wang, and Guoxian Yu. 2026. MLLM Enriched Explainable Multiple Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 28373–28381.

[53] Shan Zhang, Liangrui Ren, Jun Wang, Yanyu Xu, Carlotta Domeniconi, and Guoxian Yu. 2025. Aligning contrastive multiple clusterings with user interests. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*. 7011–7019.

[54] Yunhe Zhang, Jinyu Cai, Zhihao Wu, Pengyang Wang, and See-Kiong Ng. 2025. Mixture of Experts as Representation Learner for Deep Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 22704–22713.

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16816–16825.

[56] Taichun Zhou, Zhibin Dong, Siwei Wang, Ke Liang, Miaomiao Li, Xinwang Liu, En Zhu, and Xiangjun Dong. 2025. DPFMVC: Dynamic Progressive Fusion for Multi-view Clustering. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 1102–1111.

[57] Wenjie Zhu and Bo Peng. 2020. Sparse and low-rank regularized deep subspace clustering. *Knowledge-Based Systems* 204 (2020), 106199.

[58] Wenjie Zhu, Bo Peng, and Wei Qi Yan. 2026. Deep inductive and scalable subspace clustering via nonlocal contrastive self-distillation. *IEEE Transactions on Circuits and Systems for Video Technology* 36, 3 (2026), 3624–3637.

[59] Wenjie Zhu and Wei Qi Yan. 2026. Inductive multiple clustering based on weakly-supervised salient representation learning. *Expert Systems with Applications* 296 (2026), 129082.